Multi-facet Document Representation and Retrieval

Karam Abdulahhad, Jean-Pierre Chevallet, Catherine Berrut

Multi-facet Document Representation and Retrieval

Karam Abdulahhad*, Jean-Pierre Chevallet**, and Catherine Berrut*

* UJF-Grenoble 1, ** UPMF-Grenoble 2, LIG laboratory, MRIM group
karam.abdulahhad,jean-pierre.chevallet,catherine.berrut@imag.fr

Abstract. This paper presents our participation in ImageCLEF2011, in the two tasks: ad-hoc image-based retrieval and case-based retrieval, of the medical retrieval track.
We participated through a simple IR model based on three hypotheses: 1) the amount of overlap between a document and a query, 2) the descriptive power of an indexing element, and 3) the discriminative power of an indexing element. We used three types of indexing elements: ngrams, keywords, and concepts, for building and checking the effectiveness of multi-facet document representation.
Although of the simplicity of our model in both documents' representation and retrieval, we could obtain good results. The eighth out of 64 runs in the ad-hoc image-based retrieval task, and the fifth out of 35 runs in the case-based retrieval task.

1 Introduction

Each information retrieval (IR) model has its advantages and drawbacks. In other words, an IR model may perform well in some cases but badly in others. In general, there is no IR model performs well in all cases [9].

Our model is not an exception. Therefore, a general mechanism to evaluate the performance of an IR model and compare it with the performance of others is needed. In this context, there are many campaigns of evaluation in IR field, e.g. CLEF[1].

First, our model is a text-based retrieval and very simple model. It uses multiple types of indexing elements, e.g. ngrams, keywords, or concepts. The goals of this research are: 1) studying the performance of the model itself, using one of the indexing element types, 2) studying the effects of using multiple indexing element types at the same time on the performance.

Second, CLEF (Cross-Language Evaluation Forum) is a yearly evaluation campaign in Multilanguage information retrieval field since 2000. ImageCLEF[2] is a part of CLEF. It concerns searching medical images through documents that contain text and images [14].

[1] http://www.clef-campaign.org/
[2] http://www.imageclef.org/

This year, ImageCLEF2011[3] [12] contains four main tracks: 1) medical retrieval, 2) photo annotation, 3) plant identification, and 4) Wikipedia retrieval. Medical retrieval track contains three tasks: 1) modality classification, 2) ad-hoc image-based retrieval which is an image retrieval task using textual, image or mixed queries, and 3) case-based retrieval: in this task the documents are journal articles extracted from PubMed[4] and the queries are case descriptions.

Third, we participated in the last two tasks: ad-hoc image-based retrieval and case-based retrieval.

The test collection of this year 2011 contains, according to each task: 1) ad-hoc image-based retrieval: about 230,000 images with their text caption written in English and 30 queries written in three languages: English, French, and German. 2) case-based retrieval: about 55,000 articles written in English and 10 queries written in English.

This paper is structured as follows: Section 2 describes in details our model and the different types of indexing elements that are used. Section 3 presents some technical aspects of applying this model for ImageCLEF2011 test collection. Section 4 discusses the obtained results. We conclude in section 5.

2 The Proposed Model

2.1 Three Types of Indexing Elements

Any IR model should contain two main components: indexing function and matching function. The goal of the indexing function is to convert documents and queries from their original form to another easy to use form.

$$Index : D \cup Q \rightarrow E^* \tag{1}$$

Where
D set of documents
Q set of queries
E set of indexing elements
E^* the set of all subsets of E

Concerning indexing elements, three different types are used: ngrams (NG), keywords (K), and concepts[5] (C). Therefore, three indexing functions are exist (one for each type).

$$Index_{NG} : D \cup Q \rightarrow E^*_{NG} \tag{2}$$

$$Index_{K} : D \cup Q \rightarrow E^*_{K} \tag{3}$$

[3] http://www.imageclef.org/2011
[4] http://www.ncbi.nlm.nih.gov/pubmed/
[5] "Concepts" can be defined as "Human understandable unique abstract notions independent from any direct material support, independent from any language or information representation, and used to organize perception and knowledge" [8]. In IR domain, to achive the conceptual indexing, each concept is associated to a set of terms that describe it [3] [7].

$$Index_C : D \cup Q \rightarrow E_C^* \tag{4}$$

Where
E_{NG} set of ngrams
E_K set of keywords
E_C set of concepts

We believe that no single type of indexing elements could completely represent the content of documents and queries, because: 1) there is no perfect indexing function [3] [2] [11]. It is always an approximate function, 2) concerning concepts, the most of resources that contain concepts, e.g. UMLS[6], are incomplete [5] [6] [1], 3) each type covers an aspect of documents and queries [10]. Ngrams cover the statistical aspect, keywords cover the lexical aspect, and concepts cover the conceptual aspect.

2.2 Matching Function

Our model, as almost all models, depends on some hypotheses. Actually, it depends on the following three hypotheses:

1. The more shared elements a document and a query have, the semantically closer they are.
2. The descriptive power of an element (local weight): the more frequently an element occurs in a document, the better it describes the document [13] [3].
3. The discriminative power of an element (global weight): the less number of documents an element appears in, the more important it is [13] [3].

By taking these hypotheses into account, our model could be formulated. For any type of indexing elements the Relevance Status Value (RSV) between a document d and a query q is:

$$RSV\left(d,q\right) = \left\|d \cap q\right\| \times \left(\sum_{e \in q} \frac{N}{N_e} \times \frac{f_{d,e}}{\left\|d\right\|}\right) \tag{5}$$

Where
$d = \{e | e \in Index\left(d\right)\}$ a document
$q = \{e | e \in Index\left(q\right)\}$ a query
$\left\|d \cap q\right\| = \left\|\{e | e \in Index\left(d\right) \cap Index\left(q\right)\}\right\|$ the number of shared elements between a document d and a query q
N the number of documents in the corpus
$N_e = \left\|\{d | e \in Index\left(d\right)\}\right\|$ the number of documents that contain the element e
$f_{d,e}$ the number of occurances of an element e in a document d
$\left\|d\right\|$ the number of elements in a document d

[6] Unified Medical Language System. It is a meta-thesaurus in medical domain. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls

2.3 Matching Function According to each Type of Indexing Elements

Now after presenting our model in general, it should be instantiated according to each type of indexing elements.

Ngrams

$$RSV_{NG}(d,q) = \|d \cap q\|_{NG} \times \left(\sum_{ng \in q} \frac{N}{N_{ng}} \times \frac{f_{d,ng}}{\|d\|_{NG}} \right) \tag{6}$$

Where
$d = \{ng | ng \in Index_{NG}(d)\}$ a document
$q = \{ng | ng \in Index_{NG}(q)\}$ a query
$\|d \cap q\|_{NG} = \|\{ng | ng \in Index_{NG}(d) \cap Index_{NG}(q)\}\|$ the number of shared ngrams between a document d and a query q
N the number of documents in the corpus
$N_{ng} = \|\{d | ng \in Index_{NG}(d)\}\|$ the number of documents that contain the ngram ng
$f_{d,ng}$ the number of occurances of a ngram ng in a document d
$\|d\|_{NG}$ the number of ngrams in a document d

Keywords: We added to this instance a new component, which is the length of keyword (the number of characters). Here we supposed that the longer a keyword is, the more information it contains [15].

$$RSV_{K}(d,q) = \|d \cap q\|_{K} \times \left(\sum_{k \in q} \frac{N}{N_{k}} \times \frac{f_{d,k}}{\|d\|_{K}} \times \|k\| \right) \tag{7}$$

Where
$\|k\|$ the number of characters in a keywords k

Concepts

$$RSV_{C}(d,q) = \|d \cap q\|_{C} \times \left(\sum_{c \in q} \frac{N}{N_{c}} \times \frac{f_{d,c}}{\|d\|_{C}} \right) \tag{8}$$

2.4 The Three Types in one Model

As we said earlier, no single type of indexing elements could cover all aspects of documents and queries. Therefore, merging all types (aspects) in one model could enhance the performance of our model [9] [4] [16]. One of the merging formulas is:

$$RSV_{all}(d,q) = RSV_{NG}(d,q) + RSV_{K}(d,q) + RSV_{C}(d,q) \tag{9}$$

To increase the chance of retrieving more documents, another component could be added to the Formula (9). The component represents an expansion

of the query q. It is $\|d \cap q\|_{expan}$: the number of shared keywords between a document d and a query q after replacing each query's concept c that does not occur in d by the set of keywords that represent c[7].

$$RSV_{all}(d,q) = RSV_{NG}(d,q) + RSV_{K}(d,q) + RSV_{C}(d,q) + \|d \cap q\|_{expan} \quad (10)$$

3 Model Validation

3.1 Ad-hoc Image-Based Retrieval

In this task, image captions are used as documents and the English part of queries is just taken into account.

Text indexing: we extracted three types of indexing elements:

1. 5gram[8]: before extracting 5grams from documents and queries, we deleted all non-ASCII characters. Then we used five-characters-wide window for extracting 5grams with shifting the window one character each time.
2. Keywords: before extracting keywords from documents and queries, we deleted all non-ASCII characters. Then we eliminated the stop words and stem the remaining keywords using Porter algorithm to get finally the list of keywords that index documents and queries.
3. Concepts: before mapping the text of documents and queries to concepts, we deleted all non-ASCII characters. Then we mapped the text to UMLS's concepts using MetaMap [2].

Model variants: actually we experimented five variants of our model in this task, which are:

$$RSV_{all}(d,q) = RSV_{5G}(d,q) + RSV_{K}(d,q) + RSV_{C}(d,q) \quad (11)$$

$$RSV_{all}(d,q) = (\|d \cap q\|_{5G,K,C}) \times (sum_{5G,K,C}) \quad (12)$$

$$RSV_{all}(d,q) = RSV_{5G}(d,q) + RSV_{K}(d,q) + RSV_{C}(d,q) + \|d \cap q\|_{expan} \quad (13)$$

Where
$\|d \cap q\|_{5G,K,C} = \|d \cap q\|_{5G} + \|d \cap q\|_{K} + \|d \cap q\|_{C}$
$sum_{5G,K,C} = sum_{5G} + sum_{K} + sum_{C}$
$sum_{5G} = \left(\sum_{5g \in q} \frac{N}{N_{5g}} \times \frac{f_{d,5g}}{\|d\|_{5G}}\right)$
$sum_{K} = \left(\sum_{k \in q} \frac{N}{N_{k}} \times \frac{f_{d,k}}{\|d\|_{K}}\right)$
$sum_{C} = \left(\sum_{c \in q} \frac{N}{N_{c}} \times \frac{f_{d,c}}{\|d\|_{C}}\right)$

$$RSV_{all}(d,q) = (\|d \cap q\|_{5G,K,C}) \times (sum_{5G,K,C} + \|d \cap q\|_{expan}) \quad (14)$$

$$RSV_{all}(d,q) = ((\|d \cap q\|_{5G,K,C}) \times (sum_{5G,K,C})) + \|d \cap q\|_{expan} \quad (15)$$

[7] We supposed that each concept is represented by a set of keywords in its resource (we used UMLS as resource).

[8] 5gram is a ngram consists of five characters. We picked out 5grams because they gave the best results using ImageCLEF2010 comparing to the other ngrams.

3.2 Case-Based Retrieval

In this task, articles are used as documents. We indexed documents and queries in the same way as in the previous task. However, we extracted only two types of indexing elements (4grams, keywords) because of technical reasons.

Model variants: actually we experimented two variants of our model in this task, which are:

$$RSV_{all} = RSV_{4G} + RSV_K \tag{16}$$

$$RSV_{all}\,(d,q) = (\|d \cap q\|_{4G,K}) \times (sum_{4G,K}) \tag{17}$$

Where
$\|d \cap q\|_{4G,K} = \|d \cap q\|_{4G} + \|d \cap q\|_K$
$sum_{4G,K} = sum_{4G} + sum_K$
$sum_{4G} = \left(\sum_{4g \in q} \frac{N}{N_{4g}} \times \frac{f_{d,4g}}{\|d\|_{4G}}\right)$
$sum_K = \left(\sum_{k \in q} \frac{N}{N_k} \times \frac{f_{d,k}}{\|d\|_K}\right)$

4 Results and Discussion

Before we start representing and discussing the obtained results, we will present the names that we will use in our discussion for refering to each variant with its corresponding formula and run's name that used in the official campaign[9] (see Table 1).

Table 1. Variants' names

Varint's name	Corresponding furmula	Corresponding run's name
I_CK5G_1	Formula 11	IVSCT5G
I_CK5G_2	Formula 12	IVPCT5G
I_CK5G_Q_1	Formula 13	IVSCT5GK
I_CK5G_Q_2	Formula 14	IVPCT5GKin
I_CK5G_Q_3	Formula 15	IVPCT5GKout
C_K4G_1	Formula 16	MRIM_KJ_A_VM_Sop_T4G
C_K4G_2	Formula 17	MRIM_KJ_A_VM_Pos_T4G

4.1 Ad-hoc Image-Based Retrieval

The following table (see Table 2) contains the obtained results. The first row (Best) is the result of the first ranked run in the ad-hoc image-based retrieval task.

From one side, although of using a very simple structure (set of elements) to represent the content of documents and queries, and using a simple formula

[9] To see all results: http://www.imageclef.org/2011/medical

Table 2. The results of ad-hoc image-based retrieval task

	MAP	P@10	P@20	# rel_ret	Rank
Best	0.2172	0.3467	0.3017	1471	1
I_CK5G_1	0.2008	0.3033	0.3050	1544	8
I_CK5G_Q_1	0.2008	0.3033	0.3050	1543	9
I_CK5G_Q_2	0.1975	0.2967	0.2833	1517	10
I_CK5G_2	0.1974	0.2967	0.2833	1520	11
I_CK5G_Q_3	0.1973	0.2967	0.2833	1519	12

(see Formula 5) to compute the matching value between a document and a query, we obtained good results. The run I_CK5G_1 is ranked eighth out of 64 runs. That's because, we explicitly represented multi-facet (multi aspects) of documents and queries. In other words, three types of elements: 5Grams, Keywords, and Concepts are used and involved in the matching process.

From another side, using different fusion formulas to merge the results of using different types of indexing elements does not change a lot of things. Compare the results of the two runs: I_CK5G_1 and I_CK5G_2 (see Table 2).

In addition, the query expansion and the different formulas to integrate it into the model did not add anything. Compare the results of the three runs: $I\_CK5G\_Q\_1$, $I\_CK5G\_Q\_2$, and $I\_CK5G\_Q\_3$ with the result of the run I_CK5G_1 (see Table 2). That's because, the added value of query expansion is already compensated by using keywords and 5Grams.

Another noted result is that our model could retrieve the most relevant documents comparing to the other runs[10].

4.2 Case-Based Retrieval

The following table (see Table 3) contains the obtained results. The first row (Best) is the result of the first ranked run in the case-based retrieval task.

Table 3. The results of case-based retrieval task

	MAP	P@10	P@20	# rel_ret	Rank
Best	0.1297	0.1889	0.1500	144	1
IC_K4G_1	0.1114	0.1444	0.1444	149	5
IC_K4G_2	0.0911	0.1111	0.1278	146	10

Here also, we obtained good results. The run C_K4G_1 is ranked fifth out of 35 runs, knowing that, we used a very simple structure (set of elements), a

[10] http://www.imageclef.org/2011/medical

simple matching formula (see Formula 5), and also two simple types of indexing elements: Keywords and 4Grams. The two-facet (Keywords and 4Grams) representation of documents and queries was useful.

However, the formulas of merging the results of using different types of indexing elements were more sensitive comparing to the formulas in ad-hoc image-based retrieval task. Compare the results of the two runs: C_K4G_1 and C_K4G_2 (see Table 3). That's maybe because, we used less number of elements' types, two types (Keywords and 4Grams) in case-based retrieval comparing to three types (Concepts, Keywords, and 5Grams) in ad-hoc image-based retrieval.

5 Conclusion

We presented in this paper our approach to index and retrieve documents. We used three types of indexing elements (ngrams, keywords, concepts) for building a multi-facet document representation, and then we used a simple formula based on three hypotheses (the amount of overlap between a document and a query, the descriptive power of an indexing element, and the discriminative power of an indexing element) for retrieving documents, considering all facets (elements' types) of documents.

We obtained good results. The eighth out of 64 runs in the ad-hoc image-based retrieval task, and the fifth out of 35 runs in the case-based retrieval task, knowing that, we used a very simple structure for representing documents and queries and also a very simple ranking formula.

References

1. Karam Abdulahhad, Jean-Pierre Chevallet, and Catherine Berrut. Solving concept mismatch through bayesian framework by extending umls meta-thesaurus. In *la huitime dition de la COnfrence en Recherche d'Information et Applications (CORIA 2011), Avignon, France*, March 16–18 2011.
2. Alan R. Aronson. Metamap: Mapping text to the umls metathesaurus, 2006.
3. Mustapha Baziz. *Indexation conceptuelle guide par ontologie pour la recherche d'information.* Thse de doctorat, Universit Paul Sabatier, Toulouse, France, dcembre 2005.
4. Nicholas J. Belkin, C. Cool, W. Bruce Croft, and James P. Callan. The effect multiple query representations on information retrieval system performance. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 339–346, New York, NY, USA, 1993. ACM.
5. O Bodenreider, A Burgun, G Botti, M Fieschi, P Le Beux, and F Kohler. Evaluation of the unified medical language system as a medical knowledge source. *J Am Med Inform Assoc*, 5(1):76–87, 1998.
6. Olivier Bodenreider, Anita Burgun, and Thomas C. Rindflesch. Lexically-suggested hyponymic relations among medical terms and their representation. In *in the UMLS, in Proceedings of TIA2001, 1121*, 2001.
7. Jean-Pierre Chevallet. endognes et exognes pour une indexation conceptuelle intermdia. Mmoire d'Habilitation a Diriger des Recherches, 2009.

8. Jean-Pierre Chevallet, Joo Hwee Lim, and Thi Hoang Diem Le. Domain knowledge conceptual inter-media indexing, application to multilingual multimedia medical reports. In *ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007), Lisboa, Portugal*, November 6–9 2007.
9. W. Bruce Croft. Incorporating different search models into one document retrieval system. *SIGIR Forum*, 16:40–45, May 1981.
10. Padima Das-Gupta and Jeffrey Katzer. A study of the overlap among document representations. *SIGIR Forum*, 17:106–114, June 1983.
11. Christopher Dozier, Ravi Kondadadi, Khalid Al-Kofahi, Mark Chaudhary, and Xi S. Guo. Fast tagging of medical terms in legal text. In *ICAIL*, pages 253–260, 2007.
12. Jayashree Kalpathy-Cramer, Henning Müller, Steven Bedrick, Ivan Eggel, Alba Garcia Seco de Herrera, and Theodora Tsikrika. The clef 2011 medical image retrieval and classification tasks, 2011.
13. H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2:159–165, April 1958.
14. Henning Müller, Ivan Eggel, Joe Reisetter, Charles E. Kahn, and William Hersh. Overview of the clef 2010 medical image retrieval track, 2010.
15. Jamie Reilly and Jacob Kean. Information content and word frequency in natural language: Word length matters. *Proceedings of the National Academy of Sciences*, 108(20):E108, May 2011.
16. Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In *Text REtrieval Conference*, 1994.